Kevin Romero Peces-Barba
kromero8@ucsc.edu
1635745

CMPS 142 - Spring 2018
Homework 1 - Problem 3

Mary, the manager of a mattress store, collected a small dataset of her customer's attributes and the size of the mattress they purchased. The customers only bought two types of mattresses: King (K) or Queen (Q). Here is the data that Mary collected:

Gender	Height	Preference
M	5.2	Q
M	6.2	Q
M	6.8	Q
M	6.9	K
M	6.1	K
F	5.3	K
F	6.2	Q

Note that one of the attributes, Height, is a continuous variable. Lets assume that we will only allow binary splits for this attribute of the form Height$<$ h and Height$\geq$ h, where h lies in the dataset. However, there can be multiple such splits in one path from root to leaf.

1. Lets say Mary wants to create a decision tree to predict the mattress type that a new customer will prefer. She wants to keep 'Height' at the root node. How many possible values of h does she need to consider?

   The values Height can take are $H = \{5.2, 5.3, 6.1, 6.2, 6.8, 6.9\}$. However, taking the highest or the lowest values wouldn't be informative, because they would split the data in two sets, one containing all the data and the other one containing none.
   So the values she should consider are $H = \{5.3, 6.1, 6.2, 6.8\}$.

2. What is the entropy of labels (mattress type) in the training dataset?

$$E = \sum_{i=1}^{n} - p_i * log_2(p_i) = -\frac{4}{7} * log_2(\frac{4}{7}) - \frac{3}{7} * log_2(\frac{3}{7}) = 0.985228136$$

3. What is the optimal root node for this dataset? Show your calculations.

A. h = 5.3

This attribute splits the data into:

$h < 5.3$:

Gender	Height	Preference
M	5.2	Q

$E = 0$

$h \geq 5.3$:

Gender	Height	Preference
M	6.2	Q
M	6.8	Q
M	6.9	K
M	6.1	K
F	5.3	K
F	6.2	Q

$E = 1$

Average entropy is $E = 1$

So $IG = 0.985228136 - 1 = -0.01477$

B. h = 6.1

Splits data into:

$h < 6.1$:

Gender	Height	Preference
M	5.2	Q
F	5.3	K

$E = 1$

$h \geq 6.1$:

Gender	Height	Preference
M	6.2	Q
M	6.8	Q
M	6.9	K
M	6.1	K
F	6.2	Q

$E = 0.97095$

Average entropy is $E = \frac{2}{7}1 + \frac{5}{7}0.97095 = 0.97925$

So $IG = 0.985228136 - 0.97925 = 0.005978136$

C. h = 6.2

Splits data into:

$h < 6.2$:

Gender	Height	Preference
M	5.2	Q
M	6.1	K
F	5.3	K

$E = 0.918296$

$h \geq 6.2$:

Gender	Height	Preference
M	6.2	Q
M	6.8	Q
M	6.9	K
F	6.2	Q

$E = 0.81128$

Average entropy is $E = \frac{3}{7}0.918296 + \frac{4}{7}0.81128 = 0.913285$

So $IG = 0.985228136 - 0.913285 = 0.071943136$

D. h = 6.8

Splits data into:

$h < 6.8$:

Gender	Height	Preference
M	5.2	Q
M	6.2	Q
M	6.1	K
F	5.3	K
F	6.2	Q

$E = 0.97095$

$h \geq 6.8$:

Gender	Height	Preference
M	6.8	Q
M	6.9	K

$E = 1$

Average entropy is $E = \frac{5}{7}0.97095 + \frac{2}{7}1 = 0.97925$

So $IG = 0.985228136 - 0.97925 = 0.005978136$

E. G

Splits data into:

G = M

Gender	Height	Preference
M	5.2	Q
M	6.2	Q
M	6.8	Q
M	6.9	K
M	6.1	K

$E = 0.97095$

G = F

Gender	Height	Preference
F	5.3	K
F	6.2	Q

$E = 1$

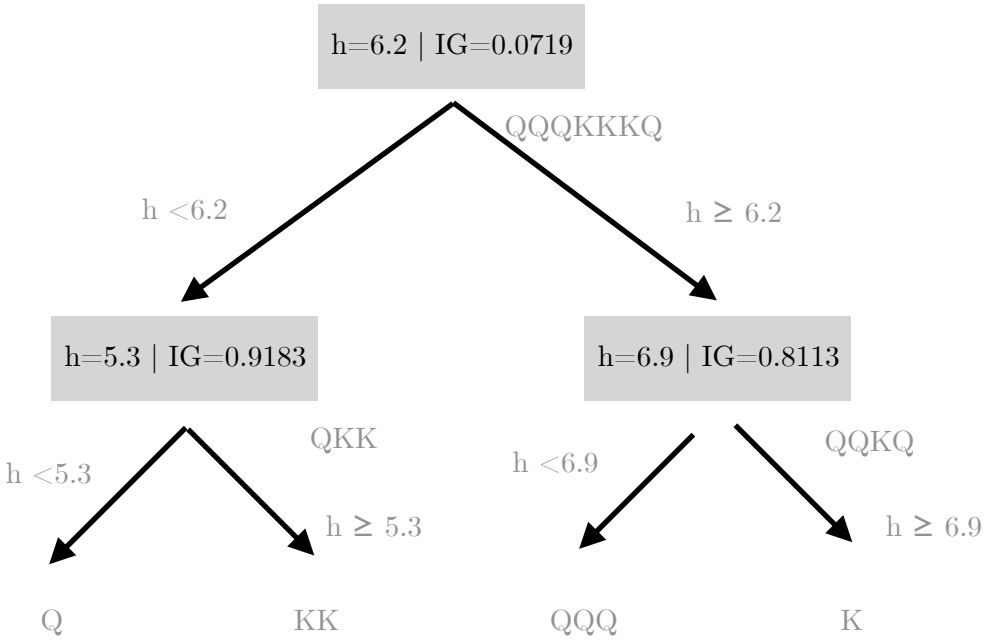
Average entropy is $E = \frac{5}{7}0.97095 + \frac{2}{7}1 = 0.97925$

So $IG = 0.985228136 - 0.97925 = 0.005978136$

Clearly, choosing C (H=6.2) has the highest Information Gain, so this would be the chosen Height as root.

4. Draw the DT that would be learned by ID3 on this dataset? Also, label each non-leaf node with the gain attained by the corresponding split. How to submit: It is okay to draw the tree by hand and include a clear picture in your pdf.

It would learn:



5. ID3 is a very popular algorithm for learning a decision tree. Does it learn an optimal tree? An optimal tree is one that has minimal depth and perfectly classifies the training data. Provide a short explanation for your answer.

No, it doesn't learn an optimal tree. It uses a greedy algorithm (selects the best attribute every iteration), and so it does not ensure to find the shortest tree because it can get stuck in local optimas.

6. Change one attribute of one example in the given dataset, so that the learned tree will contain at least one more node. As an answer to this question, provide the new training dataset (highlighting the change), and the new learned tree.

New set is:

Gender	Height	Preference
M	5.2	Q
M	6.2	Q
M	6.8	Q
M	6.9	K
M	6.1	K
F	5.3	Q
F	6.2	Q

And new tree is: